MODELING HEART DISEASE CLASSIFICATION USING ROUGH NEURAL NETWORK: A DATA-DRIVEN APPROACH TO THE CLEVELAND HEART DISEASE DATASET

Alfi Sahri [1], Nursyahrina [2], As'Ary Sahlul Irsyad[3], Nadia Aini Hafizhah[4]
[1,2,3,4]Universitas Putra Indonesia "YPTK" Padang
alfisahri18072001@gmail.com [1], nursyahrina17@gmail.com [2], ary.sahlul@gmail.com[3], nadiaaini949@gmail.com[4]

ABSTRACT

This study implements a Rough Neural Network (RNN) intelligent system method, merging rough sets with neural networks to diagnose heart disease effectively. Using the Cleveland Heart Disease Dataset, rough sets identified nine relevant features for model training, simplifying data complexity. Comparative assessment against traditional neural networks revealed the RNN model's superior performance, achieving 88.52% accuracy, 88.14% F1 score, and 88.85% AUC. This hybrid approach improves predictive accuracy while enhancing efficiency and interpretability. The findings contribute to advancing intelligent systems for heart disease diagnosis, facilitating early detection, and improving patient outcomes. Future research may explore selected features' clinical significance and RNN applicability in different contexts.

Keywords: Heart Disease Detection, Rough Neural Network, Rough Set Theory, Neural Networks, Hybrid Intelligent System

INTRODUCTION

Heart disease remains a leading cause of mortality worldwide, responsible for an estimated 17.9 million deaths annually [1]. Early and accurate detection is crucial for timely treatment and improving patient outcomes. With advances in computational methods and data availability, researchers have explored various machine learning and data mining techniques for computer-aided heart disease diagnosis [2], [3], [4], [5]. Among these, artificial neural networks (ANNs) have shown promising performance, capable of learning complex patterns and making predictions from clinical data [1], [4], [6], [7].

However, traditional ANNs often rely on all available input features, leading to increased computational complexity, overfitting, and redundancy when many features are irrelevant or correlated [8]. Feature selection or reduction techniques aim to identify the optimal subset of

relevant features, improving model efficiency, interpretability, and generalization ability [7], [8]. The rough set theory, introduced by Pawlak [5], provides a mathematical framework for feature selection by identifying reducts – minimal subsets of features that preserve the information content of the entire dataset.

The integration of rough sets with ANNs, known as Rough Neural Networks (RNNs), has been proposed as a hybrid approach leveraging the strengths of both methods. Rough sets handle feature selection and data reduction, while ANNs build predictive models from the reduced feature space [9], [10], [11]. This approach has shown potential in various applications, including consumer satisfaction prediction [11], project risk management [9], and disease diagnosis [10]. However, its application to heart disease detection still needs to be explored.

This research aims to implement an RNN model for accurate heart disease detection by combining rough sets for feature selection and neural networks for prediction. The rough set component will identify optimal reducts from clinical data, which will then be used to train an ANN model. A comparative analysis will assess the RNN's effectiveness against traditional ANNs using complete feature sets. Performance metrics, including accuracy, sensitivity, specificity, and AUC, will be evaluated, consistent with established methodologies in heart disease prediction research [2], [4], [6], [12]. The hypothesis posits that the RNN model, employing rough sets for feature selection, will deliver comparable or superior predictive accuracy alongside enhanced efficiency and interpretability. This study advances intelligent systems for heart disease diagnosis, focusing on early detection, cost reduction, and improved patient outcomes through timely interventions.

LITERATURE REVIEW

Accurate and timely detection of heart disease is crucial for effective treatment and management. Numerous studies have explored the application of intelligent systems and machine learning techniques to this problem with varying degrees of success. Traditional methods such as Support Vector Machines (SVM), Naive Bayes (NB), and Artificial Neural Networks (ANN) have shown promising results [1], [4], [7], [13]. Despite their potential, several implementations of traditional methods are often faced with difficulties due to high-dimensional data or complex feature interactions.

To address these limitations, researchers have increasingly turned to hybrid approaches that combine the strengths of multiple intelligent system methods. For instance, Akgül et al. (2020)

employed a hybrid ANN-GA (Genetic Algorithm) approach, achieving an impressive accuracy of 95.82% on the UCI Cleveland heart disease dataset. Similarly, [8] used a combination of genetic algorithms, SVM, and deep neural networks (DNN) for feature selection and classification, resulting in an AUC of 93.7% and an accuracy of 87.7%. These hybrid methods improve predictive performance and offer dimensionality reduction capabilities, reducing computational costs and enhancing model interpretability. Building upon these promising findings, a novel hybrid approach fusing Rough Set Theory (RST) and Neural Networks (NN), termed Rough Neural Network (RNN), could potentially yield even better results by leveraging the strengths of both techniques: the rule-based knowledge extraction of Rough Sets and the powerful pattern recognition capabilities of Neural Networks.

The Rough Neural Network (RNN) method combines Rough Set Theory (RST) and Backpropagation Neural Networks (BPNNs). RST preprocesses the input data to identify and retain only the core relevant features, reducing dimensionality and redundancy. This reduced feature set is then used to train the BPNN for efficient classification or prediction, leveraging the neural network's ability to handle non-linear relationships while benefiting from improved data quality, generalization, and interpretability compared to traditional neural networks [9].

METHODS

The research employs the Rough Neural Network method, combining two intelligent system methods: Rough Set and Neural Network. A Rough Set is utilized in the feature selection or data dimension reduction stage to generate a reduct, an optimal subset of features most significant and relevant in classifying patients into positive or negative classes for detecting heart disease. Meanwhile, a Neural Network generates a heart disease classification model using the reduct produced by the Rough Set method. The entire research process can be observed in Figure X.

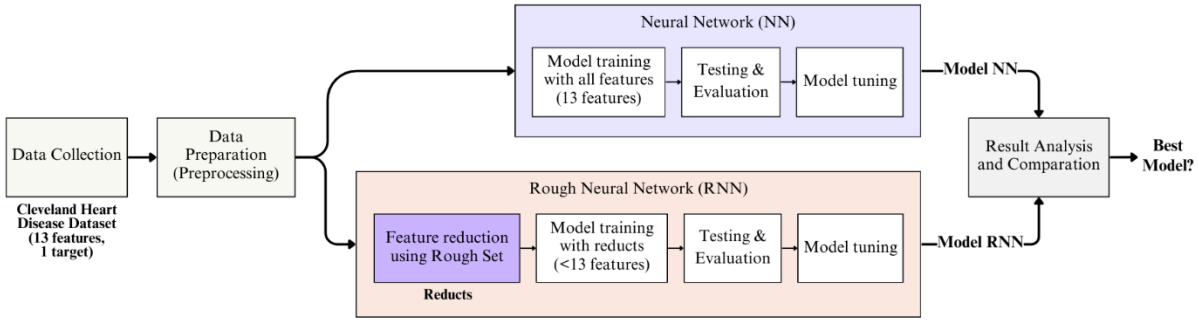

Figure 1 Research Methods

Data

The study utilized the Cleveland Heart Disease Dataset from the UCI Machine Learning Repository [14]. This dataset has undergone preprocessing procedures, comprising 303 rows and 14 columns (attributes). Each row represents patient data used to detect the likelihood of heart disease based on 13 features or variables. As a classification problem employing supervised learning, the dataset also includes one target column that classifies whether patients are indicated to have heart disease or not. A detailed description of the data is provided in Table 1.

Table 1 Dataset Description [14]

No	Attribute	Description	Data Type
1	age	Patient's age in years	Numerical
2	sex	Patient's gender	Nominal (1 = male, 0 = female)
3	cp	Type of chest pain	Nominal (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
4	treetops	Resting blood pressure (in mm Hg on admission to the hospital)	Numerical
5	chol	Serum cholesterol in mg/dl	Numerical
6	FBS	Fasting blood sugar > 120 mg/dl	Nominal (fbs > 120 = 1, fbs < 120 = 0)
7	resting	Resting electrocardiographic results	Nominal (0 = normal, 1 = ST-T wave abnormality, 2 = left ventricular hypertrophy)
8	Halacha	Maximum heart rate achieved	Numerical
9	exam	Exercise-induced angina	Nominal (1 = yes, 0 = no)
10	old peak	ST depression induced by exercise relative to rest	Numerical
11	slope	The slope of the peak exercise ST segment	Nominal (1 = upsloping, 2 = flat, 3 = downsloping)
12	ca	Number of major vessels (0-3) colored by fluoroscopy	Numerical
13	thal	Thalassemia	Nominal (3 = normal, 6 = fixed defect, 7 = reversible defect)
14	num	Diagnosis of heart disease (severity of narrowing of blood vessels)	Nominal (0 [negative] = <50% diameter narrowing, 1 [positive] = >50% diameter narrowing)

Data Collection and Preparation

The research is initiated by collecting data from the Cleveland Heart Disease Dataset in the UCI Machine Learning Repository. Subsequently, the collected data undergoes preparation to comprehend and assess the necessity for additional preprocessing before advancing to subsequent stages. This phase ensures the dataset is formatted correctly and cleaned for further analysis.

Modeling

First, the Rough Set method is applied using the ROSETTA software [9] to determine the most significant features for heart disease detection. The dataset is split into training and testing sets, with 80% allocated for training and 20% for testing, employing Python for the division. Subsequently, data normalization is conducted to ensure optimal results and integrated into a processing pipeline in Python to prevent any potential data leakage. Moving on to model construction, the neural network models are built utilizing the MLPClassifier implementation within the sci-kit-learn library for Python. The development and testing environment employed is the Google Collaboratory Notebook. Two distinct models are constructed: the NN model, utilizing all available features, and the RNN model, utilizing features reduced by applying the Rough Set method. Each model consists of one input layer corresponding to the number of features, one hidden layer, and one output layer housing two neurons for binary classification purposes. Following model construction, performance evaluation ensues using various metrics, including accuracy, precision, sensitivity, specificity, F1 score, and AUC, calculated using the elements of the confusion matrix: TP, TN, FP, and FN [7], [8]. Finally, hyperparameter tuning is carried out by leveraging a cross-validation search from the sci-kit-learn library in Python to optimize the number of hidden layer neurons and the initial learning rate for both NN and RNN models.

Result Analysis and Comparison

Following model training, evaluation, and tuning, the final NN and RNN model is selected based on the highest values of evaluation metrics, including AUC, F1 score, and accuracy, which collectively summarize the evaluation across both positive and negative classes. Subsequently, the selected RNN model and NN model are compared. This comparison aims to analyze the effectiveness of Rough Neural Network in detecting heart disease. By examining

the integration of Rough Set and Neural Network methodologies, insights are gained into how this approach enhances classification performance compared to traditional Neural Network methods alone.

RESULTS

Data Preparation Results

The data was downloaded from the UCI Machine Learning Repository and examined to identify preprocessing needs. Six rows with unknown values ('?') were found, with 4 in the 'ca' feature and 2 in the 'thal' feature. Given the nominal numeric data type, these missing values were replaced with the mode values for each feature. Furthermore, the target variable 'num' initially had five classes: 0, 1, 2, 3, and 4. However, per the data description, there should only be two classes: 0 for the absence of heart disease (negative) and 1 for the presence of heart disease (optimistic) [14]. To align with the description and facilitate understanding, as done in previous studies, the target was transformed into binary 0 (negative) and 1 (positive) class, with 'num' > 1 classified as 1 (positive). This resulted in 164 data points in class 0 and 139 in class 1. Before data reduction using the rough set method, data discretization was performed to transform continuous numeric data into a discrete form, as rough sets cannot handle continuous data well [9]. The discretization process utilized the Entropy/MDL algorithm in the ROSETTA toolkit.

Reduct

	Reduct	Support	Length
1	{sex, cp, chol, restecg, thalach, exang, oldpeak, slope, ca, thal}	15	10
2	{sex, cp, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal}	11	11
3	{sex, cp, chol, fbs, restecg, thalach, exang, oldpeak, slope, thal}	11	10
4	{sex, cp, chol, fbs, restecg, thalach, exang, slope, thal}	2	9
5	{sex, cp, chol, fbs, restecg, thalach, exang, slope, ca, thal}	1	10
6	{sex, cp, chol, restecg, thalach, exang, oldpeak, slope, thal}	8	9
7	{cp, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal}	1	10
8	{age, cp, chol, restecg, thalach, exang, oldpeak, slope, ca, thal}	2	10
9	{age, cp, chol, restecg, thalach, exang, slope, thal}	1	8
10	{sex, cp, chol, restecg, thalach, exang, slope, thal}	4	8
11	{cp, chol, restecg, thalach, exang, oldpeak, slope, thal}	2	8
12	{sex, cp, chol, thalach, exang, slope, thal}	1	7
13	{cp, chol, restecg, thalach, exang, slope, thal}	2	7
14	{sex, cp, restecg, thalach, exang, oldpeak, slope, ca, thal}	1	9
15	{cp, chol, restecg, thalach, exang, oldpeak, slope, ca, thal}	1	9
16	{cp, chol, restecg, thalach, exang, slope, ca, thal}	1	8
17	{sex, cp, fbs, restecg, thalach, exang, slope, thal}	3	8
18	{sex, cp, restecg, thalach, exang, slope, ca, thal}	2	8
19	{sex, cp, chol, fbs, thalach, exang, slope, ca, thal}	1	9
20	{sex, cp, chol, thalach, exang, oldpeak, slope, ca, thal}	1	9
21	{age, cp, chol, restecg, thalach, exang, slope, ca, thal}	1	9

Figure 2 List of Reducts Generated using ROSETTA

Feature Reduction using Rough Set

In this stage, 21 sets of reductions (reducts) are generated, detailed in Figure 2. Each reduction set consists of a subset of data features, ranging from 7 to 11 features. All these reduction sets are utilized in the feature selection process before the model training. The evaluation of model performance found that the 20th reduction set yielded the best results. This set comprises 9 features: {'sex,' 'cp,' 'chol,' 'thali,' 'exchange,' 'old peak,' 'slope,' 'ca,' 'thal'}.

Modeling and Evaluation of the Neural Network Model

Both models, Neural Network (NN) and Rough Neural Network (RNN), have undergone training, testing, and tuning processes to obtain models with optimal performance. In the initial stage, tuning the NN model resulted in the best model with an accuracy of 86.88%, F1 score of 86.67%, and AUC of 87.34%. The hyperparameters used involved the number of neurons in the hidden layer of 4, the initial learning rate of 0.054, the maximum iteration of 2000, and other parameters set to default.

Meanwhile, to select the RNN model in the initial stage, it is necessary to consider the 21 RNN models trained using different reduction sets. The evaluation results of these models show that the RNN model with the 20th reduction set performed the best with an accuracy of 88.52%, F1 score of 88.14%, and AUC of 88.85%. The evaluation performance visualization of all RNN models is presented in the graph in Figure 3, where the 20th RNN model achieved the highest scores in 4 out of 6 metrics. The NN and RNN models with the best performance were then selected for the final tuning process to produce the final NN and RNN models.

Fine-grained tuning was performed by considering the number of neurons in the hidden layer and the most suitable initial learning rate. In this stage, the 'sad' optimizer or Stochastic Gradient Descent was utilized, and the learning rate was set to the 'adaptive' mode to automatically adjust if there was stagnation in the loss reduction during training. This aims to prevent underfitting by ensuring the training continues until convergence. The 'tool' parameter or tolerance value was reduced to 1e-5, and 'n_iter_no_change' was increased to extend the training duration with more iterations (increase the patient), avoiding plateau areas and enhancing model performance. Following the tuning process, two additional models were developed for comparison, and their evaluation using test data was visualized using the ROC Curve, depicted in Figure 4.

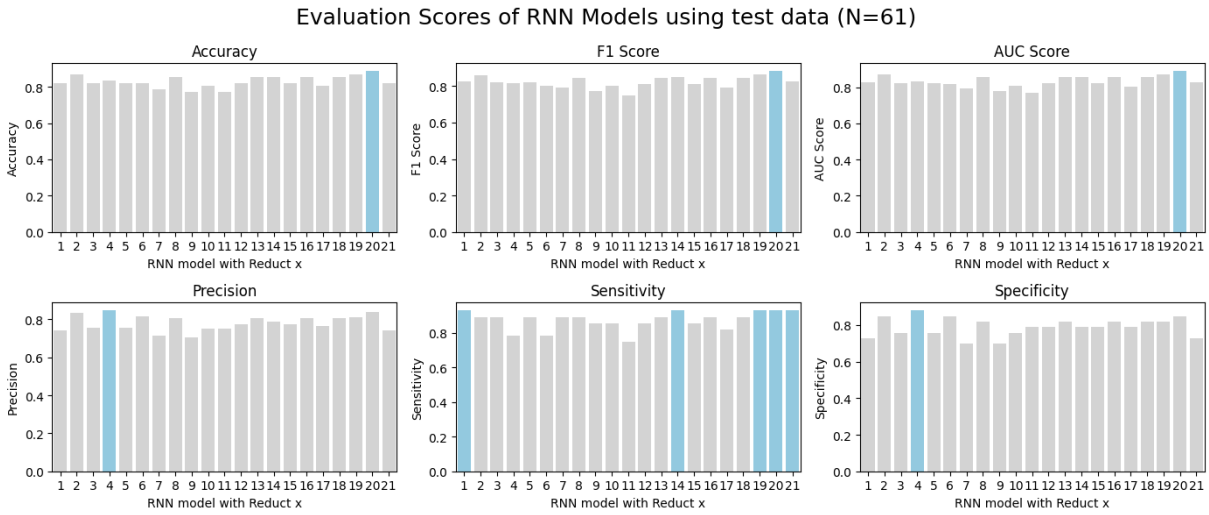

Figure 3 Evaluation Results of 21 RNN Models

Subsequently, the optimal NN and RNN models were selected for the final comparison. The chosen models are the tuned NN and the initial RNN models, selected based on their superior AUC score or the most expansive area under the ROC curve, as depicted in Figure 4. The final NN model is configured with five neurons in the hidden layer and employs an initial learning rate of 0.01. In comparison, the final RNN model consists of 7 neurons in the hidden layer and utilizes an initial learning rate of 0.056. The performance evaluation of these models is documented in Table 2.

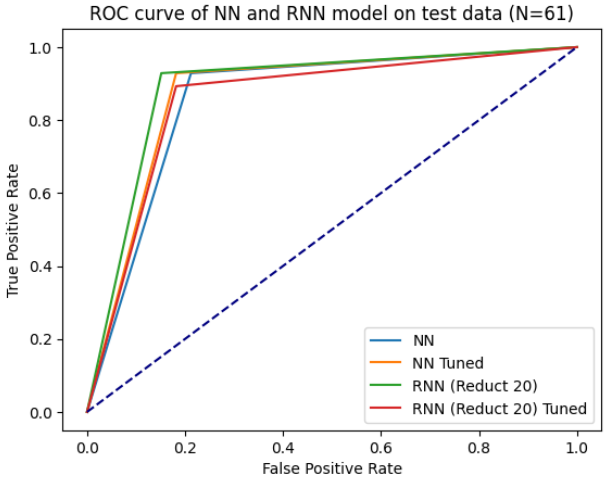

Figure 4 ROC Curve of NN and RNN Models

Table 2 Performance of the Final NN and RNN Models on the test set (N=61)

Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	AUC

NO	86.89%	81.25%	92.86%	81.82%	86.67%	87.34%
RNN	88.52%	83.87%	92.86%	84.85%	88.14%	88.85%

DISCUSSION

Based on the research findings, the hybrid Rough Neural Network (RNN) model demonstrated better performance than the traditional Neural Network (NN), which was evaluated using accuracy, F1 score, and AUC metrics. These results confirm our initial hypothesis that the hybrid method can perform comparably or even better than the single-method approach. This finding aligns with the conclusions drawn in other studies that also implemented the hybrid RNN method [5], [12]. When compared to previous research that applied NN models to the same dataset, our proposed model demonstrated superior performance, outperforming the results reported by [1], [7], [13], for instance. The rough set component successfully identified an optimal feature subset, simplifying and reducing the data dimensionality for training. Previous studies have proven this capability in other domains, such as project risk management [9]. Based on the generated reduct set, we can identify the most influential and relevant variables or parameters for detecting heart disease, which consists of 9 parameters as stated in the research results. This implies that we can achieve a reasonably accurate prediction of heart disease with only these nine parameters.

The findings highlight the effectiveness of the RNN approach in enhancing model performance while reducing computational complexity and costs associated with using an excessive number of features. By leveraging the rough set theory for feature selection and the predictive power of neural networks, the proposed model offers a promising solution for intelligent decision support systems in heart disease diagnosis. Further research could explore the interpretability and clinical relevance of the selected features and the potential for integrating additional data sources or modalities to further improve diagnostic accuracy.

CONCLUSION

This research proposed a Rough Neural Network (RNN) model that integrates rough set dimensionality reduction with neural network predictive power for accurate detection of heart disease. Using the Cleveland Heart Disease Dataset, the study implemented rough sets via ROSETTA software to identify an optimal 9-feature subset from 14 initial attributes. The RNN

model, trained on these reduced features, outperformed traditional neural networks using all features, achieving 88.52% accuracy, 88.14% F1 score, and 88.85% AUC. The superior performance of the hybrid RNN approach validates its ability to maintain predictive capabilities while improving efficiency and interpretability through relevant feature selection. This hybrid intelligent system enables early heart disease detection, reducing diagnostic costs and improving patient outcomes. Future work could explore the clinical relevance of selected features, integrating additional data sources, and extending the RNN approach to other domains.

LIMITATION

The research has limitations, including using a dataset that may cover only some relevant risk factors for heart disease and a relatively small sample size for developing neural network models. The modest number of features diminishes the advantages of feature reduction, as the traditional neural network and RNN models exhibit comparable performance. Moreover, the single hidden layer architecture may limit the model's ability to capture complex patterns optimally. While multiple evaluation metrics were employed, carefully selecting the primary metric aligned with research objectives is crucial. Future studies could investigate deeper RNN architectures and larger and more diverse heart disease datasets and extend the approach to other classification tasks to comprehensively assess its effectiveness and generalizability.

REFERENCES

H. D. Calderon-Vilca, K. E. C. Callupe, R. J. I. Aliaga, J. B. Cuba, and F. C. Mariño-Cárdenas, “Early cardiac disease detection using neural networks,” in *2019 7th International Engineering, Sciences and Technology Conference (IESTEC)*, IEEE, 2019, pp. 562–567.

J. Premsmith and H. Ketmaneechairat, “A predictive model for heart disease detection using data mining techniques,” *Journal of Advances in Information Technology*, vol. 12, no. 1, pp. 14–20, 2021, doi: 10.12720/jait.12.1.14-20.

F. F. Firdaus, H. A. Nugroho, and I. Soesanti, “A Review of Feature Selection and Classification Approaches for Heart Disease Prediction,” 2020.

N. Harika, S. R. Swamy, and Nilima, “Artificial Intelligence-Based Ensemble Model for Rapid Prediction of Heart Disease,” *SN Comput Sci*, vol. 2, no. 6, Nov. 2021, doi: 10.1007/s42979-021-00829-9.

S. Das, S. K. Pradhan, S. Mishra, S. Pradhan, and P. K. Pattnaik, "Diagnosis of cardiac problem using rough set theory and machine learning," *Indian Journal of Computer Science and Engineering*, vol. 13, no. 4, pp. 1112–1131, 2022.

S. I. Ayon, M. M. Islam, and M. R. Hossain, "Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques," *IETE J Res*, vol. 68, no. 4, pp. 2488–2507, 2022, doi: 10.1080/03772063.2020.1713916.

W. Wiharto, H. Herianto, and H. Kusnanto, "The analysis of performance model tiered artificial neural network for assessment of coronary heart disease," *International Journal of Electrical and Computer Engineering*, vol. 7, no. 4, pp. 2183–2191, 2017, doi: 10.11591/ijece.v7i4.pp2183-2191.

Wiharto, E. Suryani, S. Setyawan, and B. P. Putra, "The Cost-Based Feature Selection Model for Coronary Heart Disease Diagnosis System Using Deep Neural Network," *IEEE Access*, vol. 10, pp. 29687–29697, 2022, doi: 10.1109/ACCESS.2022.3158752.

X. Li, Q. Jiang, M. K. Hsu, and Q. Chen, "Support or risk? Software project risk assessment model based on rough set theory and backpropagation neural network," *Sustainability (Switzerland)*, vol. 11, no. 17, Sep. 2019, doi 10.3390/su11174513.

M. Z. Hasan, S. Shoumik, and N. Zahan, "Integrated use of rough sets and artificial neural network for skin cancer disease classification," in *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, IEEE, 2019, pp. 1–4.

I. A. Trivanni, "Implementasi Rough Neural Network dalam Identifikasi Kepuasan Konsumen Mediasi Bisnis," *Jurnal Kajian Ilmiah*, vol. 18, no. 2, pp. 184–194, 2018.

M. Akgül, Ö. E. Sönmez, and T. Özcan, "Diagnosis of heart disease using an intelligent method: A hybrid ANN – GA approach," in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2020, pp. 1250–1257. doi: 10.1007/978-3-030-23756-1_147.

D. Pradana, M. L. Alghifari, M. F. Juna, and D. Palaguna, "Klasifikasi Penyakit Jantung Menggunakan Metode Artificial Neural Network," *Indonesian Journal of Data and Science*, vol. 3, no. 2, pp. 55–60, 2022.

A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart Disease," *UCI Machine Learning Repository*. 1988. doi: https://doi.org/10.24432/C52P4X.